The role of preference consistency, defaults and musical expertise in users' exploration behavior in a genre exploration recommender

Yu Liang

Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
y.liang1@tue.nl

Martijn C. Willemsen

Eindhoven University of Technology
5600 MB Eindhoven
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
m.c.willemsen@tue.nl

ABSTRACT

Recommender systems are efficient at predicting users' current preferences, but how users' preferences develop over time is still under-explored. In this work, we study the development of users' musical preferences. Exploring musical preference consistency between short-term and long-term preferences in data from earlier studies, we find that users with higher musical expertise have more consistent preferences at their top-listened artists and tags than those with lower musical expertise. Users typically chose to explore genres that were close to their current preferences, and this effect was stronger for expert users. Based on these findings we conducted a user study on genre exploration to investigate (1) whether it is possible to nudge users to explore more distant genres, and (2) how users' exploration behaviors within a genre are influenced by default recommendation settings that balance personalization with genre representativeness in different ways. Our results show that users were more likely to select the more distant genres if these were presented at the top of the list. However, users with high musical expertise were less likely to do so, consistent with our earlier findings. When given a representative or mixed (balanced) default for exploration within a genre, users selected less personalized recommendation settings and explored further away from their current preferences, than with a personalized default. However, this effect was moderated by users' slider usage behaviors. Overall, our results suggest that (personalized) defaults can nudge users to explore new, more distant genres and songs. However, the effect is smaller for those with higher musical expertise levels.

CCS CONCEPTS

• **Human-centered computing** → **User studies; User models; Interactive systems and tools**; • **Information systems** → **Recommender systems**.

This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8458-2/21/09.
<https://doi.org/10.1145/3460231.3474253>

KEYWORDS

Music genre exploration, Musical expertise, Default, Nudge, Preference consistency

ACM Reference Format:

Yu Liang and Martijn C. Willemsen. 2021. The role of preference consistency, defaults and musical expertise in users' exploration behavior in a genre exploration recommender. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21)*, September 27-October 1, 2021, Amsterdam, Netherlands. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3460231.3474253>

1 INTRODUCTION AND RELATED WORK

Preferences of users are not static, but rather change over time [13]. The dynamics of user preferences has led to some discussions about the difference between users' short-term and long-term preferences in recommender systems. In the recommender system literature, short-term preferences often refer to less stable preferences that only consider a user's recent preferences, while long-term preferences are usually more stable and take into account historical data from a much longer period, such as the whole user-item interaction history [17, 22, 27]. However, little earlier work has investigated the consistency between the short-term and long-term preferences from users' behavioral data. Understanding the relation between users' short-term and long-term preferences can help to reveal how users' preferences develop over time, and further shed light on how users would behave when exploring new preferences as well as how to support users in this process with exploration tools.

1.1 The relation between short-term and long-term preferences

In the recommender system literature, different distinctions are made between short-term and long-term preferences. In session-based recommender systems, short-term preferences are often defined as users' preferences within a session, and long-term preferences are used to refer users' preferences across sessions or the whole history [25, 27]. For instance, Xiang et al. [27] proposed a graph-based recommendation approach to combine users' short-term preferences within a session and long-term preferences learned from users' whole history. Within the context of conversational recommendation, Thompson et al. [25] defined users' short-term preferences as preferences within a single conversation (session), and long-term preferences as preferences over many conversations. De Boom et al. [2] distinguished short-term preferences and long-term preferences in music recommendation: short-term

preferences modeling was more focused on the last few tracks listened, while the long-term preferences were modeled with a larger time window across sessions. Additionally, Tajala et al. [22] made a distinction between short-term and long-term preferences in the movie domain: long-term preferences are referred to user-item ratings within the whole history, modeled by a long-term preference vector [22]. Despite the distinctions made between short-term and long-term preferences, the relation between the two is under-explored. It is not yet clear how users' short-term preferences correspond to the long-term preferences, and how stable user's preferences are. In the psychological domain, Lamont and Webb [14] distinguished people's short-term and long-term musical preferences within a study to investigate what makes a person's favorite song of the day. The authors did not define short-term and long-term preferences explicitly, but it was suggested that short-term preferences were captured by songs people listened on daily basis or weekly basis, and long-term preferences were represented by songs listened a month ago or even longer. Results show that 90% of people's favorite songs of a day were heard on the day they were reported, with only a few songs from people's long-term favorites, which suggests that people's short-term preferences may not be very consistent with their long-term preferences.

In the music domain, the stability of users' musical preferences could be related to their musical expertise. Studies have shown that users with different musical expertise, as measured by their musical sophistication level [19], have different listening behaviors towards diversity [5], and also perceive or use the music recommender system differently [4, 9, 10, 12, 18]. Therefore, one would also expect that musical experts might know better what aspects of music they like and dislike, and therefore might be more consistent in their music listening behavior, showing more consistent preferences over time, than users with less musical expertise.

1.2 Preference development and exploration

To support users' exploration in the music domain and address the so-called "filter bubble" issues [21], various exploration tools have been proposed to encourage users to explore their blind spots [26], support music discovery [1, 11], guide users to explore from their current preferences [23], and support users to explore a new music genre [15, 16]. Within the context of music genre exploration, our previous work [15, 16] has also shown that users with higher musical sophistication levels for active engagement (MSAE) perceived the music genre exploration tool to be more helpful than those with lower musical sophistication levels. However, there is limited understanding about how the exploration behaviors of users, when using these tools, are related to their musical expertise and the consistency of their short and long-term preferences. For example, are users with more stable preferences less or more likely to explore away from their current preferences, and how does this relate to their musical expertise? Is the deeper knowledge of a musical expert beneficial in seeing what and what not to explore, or is it harder to actually nudge [24] an expert to explore their blind spots [26]? From a recommender perspective, exploring new preferences is a difficult task. As recommender systems are built on existing historical data of users, they are by design mostly reinforcing the current preferences [3]. Jesse and Jannach [8] argue that recommender systems

can be seen as digital nudges, as they (like regular nudges) try to help user make better decisions. Nudges, as originally intended, are not necessarily personalized and are often one-size-fits-all. They should work especially well for users who do not have strong preferences and who can use some help in decision making. However, when considered from a recommender perspective, nudges could also be quite helpful in shaping the choice environment of a recommender system. Explicit digital nudges might be particularly helpful when users have more than one goal to attain with the system [8], such as an additional goal to explore. By changing the decision structure, reordering the list of recommendations, hiding alternatives that are predicted to have lower ranking, or preselecting items by making specific items default options, one could promote items that are on the one hand sufficiently personalized to be relevant to the user, but on the other hand trigger the user to explore a bit out of her comfort zone. However, it is not clear yet to what extent users can be nudged when already being recommended relevant personalized items, and how much that depends on their preference stability and expertise.

1.3 Research goal of the current study

To understand the relation between preference development, music exploration and users' musical expertise, we first investigate the relation between users' preference consistency and their musical expertise in Sec. 2 by analyzing existing user data collected from several of our music recommendation studies [15, 16]. We expect musical experts to have more consistent preferences because they might understand their own musical preferences better and have a more specialized taste. Users' preference consistency may also influence their exploration behaviors for new preferences. Therefore in Section 3, we investigate how users' genre selection behavior is related to their preferences and their musical expertise using data from a previous study, where users were explicitly asked to explore potential new music genres [15]. Based on these analyses, we formulate more specific research questions that guide the new study, as reported in Sec. 4. In the new study, we investigate whether it is possible to nudge users to explore away from their current preferences. We nudge users on their selection of genres, by preselecting and reranking by default either the closer or more distant genre. Additionally, we also investigate how personalized users would like their recommendations to be, by providing them with an interactive exploration tool with different default levels of personalization. We further explore how the effects of these defaults differ for users with different musical expertise. We expect that the defaults for less personalization levels can promote users' exploration from users' current preferences, but to a lesser extent for users high on musical expertise, as they might have a more stable and specialized taste already, therefore are less likely to make a change.

2 MUSICAL PREFERENCE CONSISTENCY

In this analysis, we look into the relation between the preference consistency of users on Spotify and their musical expertise. The data used in this analysis was collected from two online lab studies [15, 16] and two (unpublished) real-world events. Specifically, participants needed to login with their Spotify account, and for each

participant we retrieved up to 50 top-listened¹ tracks and artists in the three time spans: short-term (last 4 weeks), medium-term (last 6 months) and long-term (across the whole listening history). Additionally, participants needed to fill in their basic demographics and we measured their musical expertise with the Goldsmiths Musical Sophistication survey on Active Engagement (MSAE) and Emotion (MSE) [19]. The participants were mainly recruited from the JFS participant pool of Eindhoven University of Technology, or via convenience sampling. Studies were all approved by the institutional ethics review board of the university and participants gave informed consent on using their data. In total, we include 319 participants (age: 24.8 ± 7.7 , 143 female, 175 male and one with other gender). The average number of top listened tracks were 49 ± 11 , 47 ± 10 , 49 ± 6 , and the average number of the top listened artists were 24 ± 17 , 42 ± 15 , 45 ± 14 for short, medium and long-term respectively. Note that for some users, the number of their top-listened tracks or artists was less than 50, probably due to limited recent interactions with their Spotify account.

2.1 Methods

The consistency between users' short, medium and long-term preferences is measured at three different levels: track-level, artist-level and tag-level. The track-level and artist-level consistency are computed with Jaccard similarity between a user's top-listened tracks/artists of the three terms. The tag-level consistency is calculated as the cosine similarity between the tags of the user's top-listened artists. Since three types of consistency can be computed for each user, a multilevel linear regression is applied to evaluate the effect of users' musical expertise on preference consistency between the 3 pairs (short versus medium-term: *sm*, short vs. long-term: *sl*, medium versus long-term: *ml*), with the *sm* consistency as the baseline. The corresponding results are presented in Table 1 and Figure 1. Only the Active Engagement (MSAE) part of the musical sophistication is used in the regression model. The factors MSE, gender and age are excluded due to non-significant effects.

2.2 Different levels of consistency

2.2.1 Track-level Consistency. The results of track-level consistency are shown in Table 1, column *Track-level* and Figure 1(a). The figure shows that the short-medium term consistency is always higher than the short-long term consistency (the negative effect of *sl* in the model), indicating that users' short-term preferences at track-level are more consistent with their medium-term than with their long-term preferences. In general, there is no difference between the short-medium term consistency and the medium-long term consistency. The short-medium and short-long term consistency do not depend on MSAE, but the medium-long term consistency decreases with MSAE (significant interaction of *MSAE:ml* in the model), indicating that medium-long term consistency is larger for low MSAE users.

2.2.2 Artist-level Consistency. One drawback of the track-level consistency is that it considers tracks from the same artist to be different, while tracks from the same artist could actually be very similar to each other. Therefore, artist-level consistency might serve

as a more accurate way for computing preference consistency. We included 305 out of 319 participants in this analysis, as 14 participants did not have top-listened artists for at least one term. The results of the multilevel linear regression model at artist-level are presented in Table 1, column *Artist-level* and Figure 1(b). Similar to the track-level consistency, the short-medium term consistency at the artist-level is higher than the short-long term consistency (negative effect of *sl*). However, at the artist-level, the medium-long term consistency is always higher than the short-medium term consistency, suggesting that users' top-listened artists in the medium-term are more consistent with the top-listened artists in the long-term than with the short-term (positive effect of *ml*). With the increase of MSAE, consistency across all three terms seems to increase (positive main effect of *MSAE*). However, this effect is strongly reduced for the medium-long term consistency (negative interaction effect of *MSAE:ml*), as reflected by the smaller slope of this line in Figure 1(b). This figure also shows the net effect: the consistency between the three terms is more equal and higher for users high on MSAE. This indicates that at the artist-level, users with higher MSAE have more stable and stronger consistency across their short, medium and long term preferences.

2.2.3 Tag-level Consistency. The results of the tag-level consistency are shown in Table 1 column *Tag-level* and Figure 1(c). We included 303 out of 319 participants in this analysis, since the other 16 participants did not have enough tags for at least one time span. Similar to the artist-level consistency, at the tag level, the medium-long term consistency is higher than the short-medium term consistency, and the short-medium term consistency is higher than the short-long term consistency. Similarly, there is an overall increase of consistency with higher MSAE, except for the medium long-term consistency which is more stable across different MSAE levels (negative interaction of *MSAE:ml*). Overall, we see that users with higher MSAE also tend to be more consistent in their preferences at the tag-level.

2.2.4 Summary. Looking at all three plots in Figure 1, we observe that preference consistency for tracks is lower than for artists and tags. Consistency goes down somewhat with MSAE at the track-level but increases with MSAE for the artist and tag level. Additionally, the short-medium term and medium-long term consistency are higher than the short-long term consistency on all measures and for all levels of MSAE. However, the differences seem to reduce with higher MSAE, especially for the artist-level and tag-level consistency, showing that users with more musical expertise (higher on MSAE) are more consistent in their preferences.

3 PREFERENCE CONSISTENCY AND GENRE EXPLORATION SELECTION

As users with high MSAE seem to have more consistent preferences over time, they might know better what they like and might not change their listening behavior much over time. This also suggests that it might be harder for users high on MSAE to move away from their current preferences by means of novel recommendations. In this section, we investigate how the genre selection behavior differs for users with different levels of MSAE when exploring new music genres, with data collected from a previous study [15].

¹This number is restricted by Spotify API (<https://developer.spotify.com/documentation/web-api/reference/>).

Figure 1: The relation between MSAE and predicted consistency from the multilevel regression model. The shaded area represents the 95% confidence interval. The scatter plot represents the original data distribution (slightly jittered for the display purpose).

Table 1: Preference consistency as modeled on the track, artist and tag level

	Track-level	Artist-level	Tag-level
MSAE (centered)	-0.008 (0.012)	0.025 (0.011)*	0.054 (0.010)***
sl	-0.126 (0.011)***	-0.101 (0.008)***	-0.070 (0.009)***
ml	-0.010 (0.011)	0.089 (0.008)***	0.083 (0.009)***
MSAE (centered): sl	-0.004 (0.011)	0.001 (0.008)	0.011 (0.009)
MSAE (centered): ml	-0.028 (0.011)***	-0.018 (0.008)***	-0.037 (0.009)***
Constant	0.237 (0.013)***	0.332 (0.012)***	0.796 (0.011)***
N	957	915	909
AIC	-406.20	-772.59	-809.88
BIC	-367.29	-734.04	-771.38
$R^2_{GLMM(m)}$	0.066	0.130	0.164
Random Effect			
# of Participants	319	305	303
Participant Standard deviation	0.188	0.181	0.148

Note. We set short-medium term preference consistency (sm) as the baseline, so the intercept represents the sm consistency. sl = short-long term preference consistency; ml = medium-long term preference consistency (both relative to sm, the baseline). AIC: Akaike information criterion; BIC: Bayesian information criterion. The standard error is shown in brackets. *** $p < .001$; ** $p < .01$; * $p < .05$.

3.1 The previous study

In the analysis of the previous study on music genre exploration [15], we mostly focused on users' perception about the recommendations, however in this section, we focus on an new additional analysis of users' genre selection behavior before receiving the recommendations. We only briefly discuss the study procedure here and we recommend readers to refer to the previous work for more details. During the study, participants needed to first login with their Spotify account and fill in a survey on their Musical Sophistication Index [19] on Active Engagement (MSAE) and Emotion (MSE). They were then asked to select a music genre from a genre list to explore. The music genre list contains 10 different music genres from Allmusic.com²: Avant-garde, Blues, Classical, Country, Electronic, Folk, Jazz, New-age, Rap and R&B (ordered alphabetically), with tracks of each genre scraped from Spotify Web API. After selecting a genre, participants subsequently received the recommended tracks from the recommendation algorithms.

²<https://www.allmusic.com/genres>

3.2 Genre selection behaviors

In the analysis, we include 130 (out of 156) responses of participants with enough activities on Spotify, i.e., the participants should have top-listened artists, tracks and tags in all the three terms (average age: 22.7 ± 4.9 , 62 female, 68 male). To investigate the relation between genre selection and user preference, we first computed how close each genre was to the user's short, medium and long-term preferences. The closeness is calculated as the tag-level similarity, i.e., cosine similarity between the user's top-listened tags and the genre profile represented by the tags of all artists from that genre. We use the tag-level similarity rather than track/artist-level similarity, as the tag-level similarity can relieve the data sparsity issue caused by the potential gap between the selected genre and the user's current preferences. As users needed to select one music genre from a list of 10 different genres to explore, we applied a conditional logistic regression³ to test whether users are more likely to select genres close to their short-term, medium-term and long-term preferences, and how this differs for users with different musical

³Conditional Logistic regression is a regression technique that models a choice of one out of N ($N > 2$) items

Figure 2: Selected genre closeness for users with low or high MSAE/MSE to their short-term, medium-term and long-term preferences. The numbers are the means. The error bar represents 1 standard error of the mean.

expertise (MSAE and MSE). We found that users are more likely to select genres to explore that are close to their preferences for all terms: short-term ($\beta = 4.033, p < .001$), medium-term ($\beta = 3.919, p < .001$) and long-term ($\beta = 3.839, p < .001$). Additionally, we notice a positive interaction effect of MSAE/MSE with the short-term closeness on the genre selection: with the increase of MSAE ($\beta = 1.841, p < .05$) or MSE ($\beta = 2.811, p < .05$), users are more likely to select a genre close to their short-term preferences. Note that we notice no significant effect of alphabetical ordering on genre selection.

To better understand the results of the conditional logistic regression, users are split into low or high MSAE/MSE based on the median-split of the continuous MSAE and MSE scale, and their closeness to the selected genre is plotted in Figure 2. We see that for users with high MSAE/MSE, the selected genre is almost equally close to their preferences in all terms, and the somewhat higher closeness of the medium term is not significant in the model. For users with low MSAE/MSE, we see on overall that genre closeness is lower and there are larger differences between users' short, medium and long-term preferences. The results suggest that users tend to explore from the (well-developed) long-term preferences. Users high on MSAE have more stable preferences, so the selected genre is equally close to their short, medium or long-term preferences. While users low on MSAE have less stable preferences, thus they tend to explore genres closer to the well-developed long-term preferences than their short-term preferences. Overall, we observe that users high on musical expertise clearly select genres closer to their current (short, medium and long-term) preferences than those with low musical expertise.

4 CURRENT STUDY: THE EFFECTS OF DEFAULTS IN MUSIC GENRE EXPLORATION

Results from Section 3 suggest that users tend to explore genres that are close to their current preferences. This effect applies to all users, but is stronger for users that are high on musical expertise, due to their higher preference stability. For genre exploration to be effective, it might be beneficial to nudge users to explore genres that are somewhat further away from their current preferences.

Therefore, we would like to investigate whether a default order of genres that promotes more distant genres (with the most distant genre pre-selected) is efficient to nudge users to explore. According to Jesse and Jannach [8], changing the decision structure, by pre-selecting a certain option as default or change the order of recommendations can serve as an efficient way to nudge users to explore, as the order would make items seen first (the more distant genres) to be more important to users. Moreover, since users with higher musical expertise tend to choose closer genres to explore, we also explore whether the effect of nudging differs for users with different musical expertise. Would nudges be more beneficial especially for those with high musical expertise or would they instead be less affected by a nudge?

The genre exploration tool from the previous work [15] generated recommendations at three levels of personalization. Within the selected genre, the recommended playlists were either very personalized (tracks from the genre very close to the user's current tastes), very genre representative (tracks typical of the genre but not personalized to the user's current tastes) or mixed, with tracks that are both somewhat representative and somewhat personalized. However, users were not allowed to set the preferred level of personalization themselves. In this current study, users can interactively set the personalization level of the recommendations themselves, using a slider ranging from very genre representative to very personalized. Specifically, we explore how much the preferred level of personalization within the selected genre is affected by default personalization settings, corresponding to the three slider positions (the representative, the mixed or the personalized position) and the distance between the selected genre and user's current preferences.

Our research questions are: **RQ1:** Does promoting the more distant genre through a presentation default make users explore genres further away from their current preferences? **RQ2:** How personalized do users like their playlists to be during genre exploration, and does the preferred level of personalization depend on the default slider settings and the genre distance from their current preferences? **RQ3:** Does musical expertise moderate how far users are willing to explore, in terms of both genre selection (RQ1) and the preferred level of personalization they set to the playlist during genre exploration (RQ2)?

4.1 Recommendation approaches

Based on users' feedback about missing (popular) genres in the previous study, the original dataset for music genre exploration (as described in Section 3.1) was enriched from 10 to 13 music genres following the categorization of Allmusic.com. The three new added genres are Latin, Pop/Rock and Reggae. The new dataset was constructed in the same way as described in [15]. In short, the most typical artists of each genre were first retrieved from Allmusic.com (we rely on the editors' opinion for this), and then extended this with data from the Spotify API for relevant track and artist information.

4.1.1 Genre suggestion algorithm. We use the Personalized PageRank algorithm [6, 7] to generate genre suggestions based on tags. PageRank is an algorithm used by Google to compute the likelihood of a random surfer arriving at certain web page by clicking links. It measures the importance of nodes within a graph based on the number of incoming edges and node weight. In the normal PageRank algorithm, the node weight is evenly distributed to all nodes, i.e., each node has a same weight. While in Personalized PageRank, different weights are assigned to different nodes, which allows the algorithm to prioritize nodes with higher weight. In the recommendation scenario, the graph can be a user-item graph connected by the positive user-item ratings, with higher weights assigned to nodes that represent users' current preferences, such as items positively rated by the user [20], and the relevance score of a certain item can then be calculated as the likelihood from the source nodes (with weight) to the item node. In our case, there is no user-item interaction data available, however, the links between music tags allow us to calculate the relevance scores between a user's current preferences (represented by the top-listened tags) and a certain genre (represented by the tags associated with the genre). To do so, a tag network is built as the first step. Specifically, in the genre dataset, each artist is associated with a list of tags. Given a set of artists and tags associated with the artists, an edge is created between two tags if they are associated with a same artist, and the weight of the edge increases with the number of artists associated with both tags. Then we assigned weights to each tag that appears in the user's top listened artists, based on the number of times the tag appears in the user's top artists. The relevance score between a genre and a user's current preferences (represented by the weighted node) is then calculated as the likelihood from the weighted nodes to the tags associated with the genre. The genre with the highest relevance score is returned as the closest genre.

4.1.2 Genre exploration algorithm. The genre exploration algorithm is adapted from the existing music genre exploration tool [15]. Firstly, users' current preferences are modeled by the audio features of their top listened tracks (retrieved from Spotify), with a Gaussian Mixture Model built for each audio feature dimension (*energy*, *valence*, *acousticness* and *danceability*). The personalized algorithm recommends the top tracks within the genre that match the user's current preferences in the audio feature space. Specifically, the algorithm first calculates the scores for the candidate tracks in each feature dimension, then aggregate the track rankings of all feature dimensions and return the top tracks with the highest matching scores (lowest average rankings). In a similar way, the non-personalized algorithm recommends the top

tracks matching the genre-typical tastes in the audio feature space. To enable control over personalization and representativeness, the recommendation scores from the personalized and the non-personalized are combined based on personalized weight set by the user: $score = w \times (|R| - r_{pers}(i) + 1) + (1 - w) \times (|R| - r_{repr}(i) + 1)$, in which w denotes the personalized weight the user set, R is the number of tracks within the genre, $r_{pers}(i)$ denotes the ranking of track i in the personalized algorithm and $r_{repr}(i)$ denotes the ranking of track i in the non-personalized algorithm.

4.2 Study design

4.2.1 User interface. Figure 3 depicts the genre selection interface for the two different default pre-selection and genre-closeness orders. Figure 4 depicts the interface for exploring the selected genre. The interface consists of the following elements: (a) the top 10 recommendations in a in playlist style, (b) the *trade-off* slider for adjusting the preferred level of personalization, and (c) a contour plot visualization. The contour plot is adapted from our previous work [16], where it was successful in improving the understandability and helpfulness of the genre exploration tool. The contour plot serves as a visual explanation of how the recommendations relate to the users current preferences and the music genre. Specifically, users can inspect the generated recommendations (in blue dots) with both their current preferences (represented in the orange color) and the genre typical tastes (represented in the green color) on the four audio feature dimensions used for recommendation: valence, energy, acousticness and danceability.

During the exploration, users can adjust the preferred level of personalization with the *trade-off* slider. After adjusting the slider value, both the top 10 recommendations and the blue dots in the contour plot update accordingly. Additionally, when hovering over an individual track in the recommendation list, the corresponding track is also highlighted in the contour plot, and vice versa. Users can also hover over the question mark at the top right corner of the slider and contour plot components for more explanations about the interface. Users can also save the generated playlists to their Spotify account by clicking the button *Save playlist to my Spotify* below the playlist.

4.2.2 Study conditions. A 2 x 3 between-subject design was conducted to investigate the effect of different defaults. Users were randomly assigned to one of these six conditions. The genre presentation order (with the preselected genre) was manipulated on two levels (as shown in Figure 3). The genre suggestions were either presented in increasing order of closeness with the most distant genre preselected (Figure 3(a)) or in decreasing order of closeness with the second closest genre preselected (Figure 3(b)). Note that the closest genre was not included in the list but was reported to the user: *Your Spotify profile shows that you like the genre ... most*, as shown in Figure 3. This provides users with a sense of personalization, showing that the recommender is able to predict their favorite genre, and makes sure that users at least explore one step away from their current preferences. The position of the slider was manipulated at 3 levels: the slider was either set at the fully representative position, at the fully personalized position, or in the middle.

Figure 3: Genre selection Interface

Figure 4: Interface elements: (a) the top 10 recommendations, (b) the *trade-off* slider for adjusting the personalization levels, (c) the contour plot visualization for the four audio feature dimensions: valence (negative–positive), energy (calming–exciting), acousticness (less–more acoustic), and danceability (less–more danceable). The orange contour plot represents the user’s taste, the green contour plot represents the selected genre, and the blue dots represent the recommendations.

4.2.3 *Measurements.* As our goal is to investigate the role of different defaults in users’ exploration behavior during music genre

exploration, we focus on the objective measures of users’ behavior. Users’ genre selection behavior was recorded to evaluate the

effects of different genre suggestions for nudging them to explore a distant genre. Furthermore, we measure the personalization level users set to the playlist (the final recommendation list after exploring with the slider) to test the effects of the default settings and the user-genre distance on the preferred level of personalization. As the preferred level of personalization can only show how far a users would like to move away from their current preferences within the selected genre relatively, the absolute distance between the user's current preferences and the final recommendations was also measured as an additional metric. Furthermore, as we are interested in how users' musical expertise moderates the effects of different defaults, we also measured users' musical expertise with the Goldsmiths Musical Sophistication survey on Active Engagement (MSAE) and Emotion (MSE) [19], as in the previous studies. We also recorded several other types of objective behavior of the participants, including their interaction time with the track, track interaction times, playlist saving behaviors as well as the slider usage during genre exploration. In addition to the objective measures, users' perceived control and understandability over the recommendations, as well as the playlist they created for perceived representativeness, perceived personalization, and helpfulness were measured with a post-task questionnaire. However, due to space constraints, we focus our discussion on the measures that allow us to investigate the effect of defaults on users' objective exploration behavior, and leave the analysis of other measures such as the playlist saving behavior and the subjective perceptions for the future work.

4.2.4 Participants. The participants were recruited from Prolific. We first ran a short screen survey to find target participants with an active Spotify account. As reimbursement, we paid £0.1 for each participant in the screen survey and £1.25 for each participant in the main study (with a duration of 10 minutes). The study was approved by the institutional ethics review board of Eindhoven University of Technology. In total, we received 353 valid responses from the main study (35 participants were removed from the analysis because they did not listen to any of the recommended tracks). The participants have an average age of 26.6 ($SD = 7.8$; range from 18–60), with 189 males and 164 females. The genre selection frequency is presented in Figure 5.

4.3 Study procedure

After agreeing with the informed consent, participants logged in with their Spotify account and gave permissions to access their Spotify top-listened tracks and artists. They were then asked to fill in the musical sophistication survey on Active Engagement and Emotion. After that, participants were redirected to the genre selection page, in which they needed to select a music genre to explore from a list of genres (see Fig. 3). Then, they arrived at the genre exploration page (as shown in Figure 4) and received a guided tour, which explained them the interface elements. In the task instruction, participants were given a task described as “to create a good playlist that is helpful for you to explore the genre ... using the slider”. They could make use of the visualization to see how the generated recommendations are related to the selected genre and their own preferences when using the slider. Participants could also save the recommendations to their Spotify account if they like

it. After the task, they were asked to fill in a post-task questionnaire about their experience with the system and the playlist they created.

4.4 The effects of genre suggestion

We applied conditional logistic regression to investigate the genre selection behavior of users. Genres are distinguished based on the relevance score from the Personalized Pagerank algorithm (see Sec. 4.1.1). The top six genres with the highest relevance scores (from the 2nd to the 7th position) are grouped as “close genres”, with remaining six genres ranking at the bottom (from the 8th to the 13th position) as “distant genres”. For the conditional logistic regression, we test whether users are less likely to select “close genres” to explore when the genre list is presented with the most distant genre first than the closest genre first (RQ1), and whether users' MSAE influence the selection of close genres (RQ3). The results show that users in general tend to select “close genres” to explore ($\beta = 1.3777, p < .001$), accompanied by a negative interaction effect between “close genres” and the presentation of the most distant genre first ($\beta = -1.5772, p < .001$), which indicates that users are less likely to select from “close genres” to explore when the most distant genres are presented first. Moreover, a positive interaction effect of “close genres” with the default for more distant genres and MSAE ($\beta = 0.3744, p < .05$) suggests that with increasing MSAE scores, users are more likely to select from “close genres” to explore even when the default is set for exploring distant genres. Overall, we found that users can be nudged to explore a more distant genre when distant genres are presented first. However, the effect of nudging differs for users with different MSAE scores: users with higher MSAE are less sensitive to the default for promoting distant genres. The results are consistent with our previous findings on preference consistency: with the increase of MSAE, users are more stable in their preferences, therefore less likely to move away from their current preferences.

4.5 The effects of default slider positions

4.5.1 The final slider position. We applied a linear regression on final slider position to test the effects of the defaults on the level of personalization, with the fully personalized default as the baseline. We see that users are influenced by the default settings in general. Compared to the personalized default (intercept=0.65, $p < .001$), users with the mixed default ended up with a somewhat lower (less personalized) position on the scale ($\beta = -0.09, p = .05$). Users with the representative default set their final slider at a substantially less personalized level ($\beta = -0.16, p < .001$). Overall, the mixed and representative defaults help users to set the slider at a less personalized level, which should help them to explore more genre-specific tracks. Notably, we observe no effect of the musical expertise on the preferred level of personalization that users set to the playlist. Additionally, we evaluate whether the final slider position is influenced by the distance between the user and the genre. The user-genre distance is measured as the euclidean distance between the averaged feature vector of users' top tracks and the averaged feature vector of all tracks in the selected genre in the 4-dimensional audio feature space (energy, valence, acousticness and danceability), as this is the space that users have control over the recommendations and set preferred level of personalization.

Figure 5: Genre selection frequencies

Unexpectedly, we found no effects of user-genre distance on the preferred level of personalization.

Additionally, we observe different types of slider usage, depending on how much users interacted with the slider. Based on how many times users used the sliders, users are split into three groups: low slider interaction group (slider interactions ≤ 1 , $N = 96$), medium slider interaction group ($1 < \text{slider interactions} \leq 4$, $N = 106$), and high slider interaction group (> 4 slider interactions, $N = 151$). Inspecting the interaction effects of the default slider position and slider interaction times on the final slider position using a linear regression model, we observe that the slider interaction times moderate the effect of the default slider position on the final slider position, as presented in Figure 6. Users with low slider interaction times (the “Low” group in Figure 6) were strongly affected by the default slider settings: they stay close to the default setting given the representative default. However, users with medium slider interactions (the “Middle” group) behaved in an opposite way: they moved the slider to the other side of the default position given either the representative or personalized default (reflected by a positive interaction effect of the representative default and medium slider interactions: $\beta = 0.523$, $p < .001$). Finally, users with high slider interaction times (the “High” group in Figure 6) tend to end up close to the middle (mixed) position, and are less affected by the default (reflected by a positive interaction of the representative default and high slider interaction: $\beta = 0.354$, $p < .001$). For the mixed default, no difference of the last slider position among users with different interaction times, as also reflected by the non-significant interactions in the model. For the personalized default, users in general stay close to the personalized default. However, users with medium slider interactions tend to move the slider away from the personalized position (reflected by a negative interaction effect of the personalized default and medium slider interactions: $\beta = -0.307$, $p < .001$). Across all conditions, we see that the personalization slider for exploration has value for the users. Users typically do not choose a fully personalized playlist or very representative playlist, but they seem to explore different slider positions and on average find the best position to be a trade-off

Figure 6: The estimated final slider position across the three default settings for users with different slider interaction times. Slider position at 1.0 indicates fully personalization. Error bars indicate 1 standard error.

between personalization and genre representativeness. Therefore, the next question is whether these effects also influence how much they actually explore away from their current preferences.

4.5.2 The absolute distance explored from the current preferences.

To evaluate the effectiveness of the default settings on genre exploration, we measured the absolute distance between the created playlist and users’ current preferences. The absolute distance is calculated as the euclidean distance between the average audio feature vector of users’ top tracks and the average audio feature vector of tracks in the playlist. A path model (see Figure 7) is built to illustrate how different study conditions influence the absolute distance users explored from the current preferences, and how this effect is mediated by the last slider position as well as the user-genre distance. From the path model, we observe that the effect of defaults on the absolute distance that users explore from their current preference is affected by two different pathways: one via their last slider position, and the other via the user-genre distance. The effect of last slider position is negative ($\beta = -.380$), showing that more personalized slider positions reduce the absolute distance. The slider at a more personalized position makes the recommendations closer to users’ current preference, thus reduces the absolute distance. Consistent with the findings in Sec. 4.5.1, the mixed ($\beta = -.092$) and representative ($\beta = -.166$) default slider settings make users set the final slider at a less personalized level than the personalized default. These results indicate that these defaults in the end should have an overall positive effect on the absolute distance that users explore. Indeed, inspecting the total effect of the defaults on the absolute distance via last slider position, we notice that the representative default positively influences the absolute distance users move away from their current preferences ($\beta = .063$, $se = .018$, $p < .001$). The mixed default also seems to make users move further away from their current preferences ($\beta = .035$, $se = .018$, $p = .051$), however this effect is smaller and only marginal significant. The another pathway due to the genre selection phase, shows a positive effect of the default genre suggestion order on the user-genre distance ($\beta = .125$), i.e., the default for distant genres makes users explore genres that are further away, and the user-genre distance further

Figure 7: The path model for the effect of conditions on the absolute euclidean distance between the created playlist and users' current preferences. Significant levels: * $p < .001$; ** $p < .01$; * $p < .05$. Numbers on the arrows indicate standardized coefficients and standard error (in brackets) of the effect**

positively influences the absolute distance users explored from their current preference. The total effect of the default genre suggestion order on absolute distance via user-genre distance is positive and significant ($\beta = .084, se = .018, p < .001$). This indicates that the distant genre default makes users select from more distant genres to explore, and subsequently lead them to explore further away from the current preferences than the default for close genres. Overall, the path model shows that default settings in both the genre selection phase and for the trade-off slider have a strong influence on how far away users explore from the current preferences.

5 CONCLUSION AND FUTURE WORK

Overall, our results show that both the defaults and users' musical expertise play a role in users' exploration behaviors during music genre exploration. Exploring the musical preference consistency between the short-term and long-term preferences in user data from earlier studies, we found that users with higher MSAE are more consistent at their top-listened artists and tags than those with lower musical expertise, which indicates that users with high MSAE are more stable in their preferences and thus could be more conservative when exploring potential new preferences. This was supported by the subsequent findings that although users in general tend to select a close genre to explore, users with high musical expertise are even more likely to explore genres close to their preferences than users with low musical expertise. In the new study presented in Sec. 4, we further investigate the effects of default and musical expertise on users exploration behavior. The results show that users are more likely to select "distant genres" to explore when the genres are presented in an order of more distant genres first. However, the effect of default for distant genres on nudging users to explore "distant genres" becomes smaller with the increase of MSAE, i.e., users with higher musical expertise are less sensitive to the default for "distant genres". The default personalization level positively influences the preferred personalization level users set to the playlist for genre exploration. Overall, users tend to set the slider at a less personalized level given the mixed or the representative default than the personalized default. However, this effect is moderated by the slider interaction times of users. The effects of the defaults are further explored by looking at the absolute distance users explored from their current preferences. Compared to the personalized default, both the mixed and the representative default

nudge users to explore further away from their current preferences by making them set the slider at less personalized levels. Additionally, the default for "distant genres" also nudge users to explore further away from the current preferences by making them select "distant genres". Our results show that personalized digital nudge that promoting distant items through ordering, preselecting and setting defaults for exploration, is useful to trigger users' exploration from their comfort zone, however, the effect of nudging is moderated by users' domain knowledge to some extent, in our case the musical expertise of users. We also identify several limitations of this work. Firstly, although users' subjective perception was measured in the study, our discussion is mainly focused on users' objective exploration behavior, and how their behavior is influenced by their musical expertise and different defaults. In the future work, we will further investigate the results from the subjective measures to see whether the defaults also influence users' perceived helpfulness. Secondly, a lack of subsequent users actions makes it hard to know whether users really change their preferences with this tool or not. Therefore, we also plan to do a longitudinal study for measuring the actual preference change.

REFERENCES

- [1] Wanling Cai, Yucheng Jin, and Li Chen. 2021. *Critiquing for Music Exploration in Conversational Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 480–490. <https://doi.org/10.1145/3397481.3450657>
- [2] Cedric De Boom, Rohan Agrawal, Samantha Hansen, Esh Kumar, Romain Yon, Ching Wei Chen, Thomas Demeester, and Bart Dhoedt. 2018. Large-scale user modeling with recurrent neural networks for music discovery on multiple time scales. *Multimedia Tools and Applications* 77, 12 (2018), 15385–15407. <https://doi.org/10.1007/s11042-017-5121-z> arXiv:arXiv:1708.06520v1
- [3] Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations through Listening to Users. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) (RecSys '16). Association for Computing Machinery, New York, NY, USA, 221–224. <https://doi.org/10.1145/2959100.2959179>
- [4] Bruce Ferwerda, Mark P. Graus, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2017. How item discovery enabled by diversity leads to increased recommendation list attractiveness. In *Proceedings of the Symposium on Applied Computing - SAC '17*, Vol. Part F1280. ACM Press, New York, New York, USA, 1693–1696. <https://doi.org/10.1145/3019612.3019899>
- [5] Bruce Ferwerda and Marko Tkalcic. 2019. Exploring Online Music Listening Behaviors of Musically Sophisticated Users. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP'19 Adjunct*. ACM Press, New York, New York, USA, 33–37. <https://doi.org/10.1145/3314183.3324974>
- [6] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: The Who to Follow service at Twitter. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web* (2013), 505–514.

- [7] T.H. Haveliwala. 2003. Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering* 15, 4 (2003), 784–796. <https://doi.org/10.1109/TKDE.2003.1208999>
- [8] Mathias Jesse and Dietmar Jannach. 2021. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* 3 (2021), 100052. <https://doi.org/10.1016/j.chbr.2020.100052>
- [9] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of Individual Traits on Diversity-Aware Music Recommender User Interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) (UMAP '18). Association for Computing Machinery, New York, NY, USA, 291–299. <https://doi.org/10.1145/3209219.3209225>
- [10] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of Personal Characteristics on Music Recommender Systems with Different Levels of Controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) (RecSys '18). Association for Computing Machinery, New York, NY, USA, 13–21. <https://doi.org/10.1145/3240323.3240358>
- [11] Mohsen Kamalzadeh, Christoph Kralj, Torsten Möller, and Michael Sedlmair. 2016. TagFlip: Active Mobile Music Discovery with Social Tags. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) (IUI '16). Association for Computing Machinery, New York, NY, USA, 19–30. <https://doi.org/10.1145/2856767.2856780>
- [12] Iman Kamehkhosh, Geoffroy Bonnin, and Dietmar Jannach. 2020. Effects of recommendations on the playlist creation behavior of users. *User Modeling and User-Adapted Interaction* 30, 2 (apr 2020), 285–322. <https://doi.org/10.1007/s11257-019-09237-4>
- [13] Yehuda Koren. 2010. Collaborative filtering with temporal dynamics. *Commun. ACM* 53, 4 (apr 2010), 89. <https://doi.org/10.1145/1721654.1721677>
- [14] Alexandra Lamont and Rebecca Webb. 2010. Short- and long-term musical preferences: what makes a favourite piece of music? *Psychology of Music* 38, 2 (apr 2010), 222–241. <https://doi.org/10.1177/0305735609339471>
- [15] Yu Liang and Martijn C. Willemsen. 2019. Personalized Recommendations for Music Genre Exploration. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) (UMAP '19). Association for Computing Machinery, New York, NY, USA, 276–284. <https://doi.org/10.1145/3320435.3320455>
- [16] Yu Liang and Martijn C. Willemsen. 2021. Interactive Music Genre Exploration with Visualization and Mood Control. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 175–185. <https://doi.org/10.1145/3397481.3450700>
- [17] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies* 121 (jan 2019), 21–41. <https://doi.org/10.1016/j.ijhcs.2018.05.002>
- [18] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) (UMAP '20). Association for Computing Machinery, New York, NY, USA, 173–182. <https://doi.org/10.1145/3340631.3394844>
- [19] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. 2014. The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE* 9, 2 (feb 2014), e89642. <https://doi.org/10.1371/journal.pone.0089642>
- [20] Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2017. Tuning personalized pagerank for semantics-aware recommendations based on linked open data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10249 LNCS (2017), 169–183. https://doi.org/10.1007/978-3-319-58068-5_11
- [21] Eli Pariser. 2012. *FILTER BUBBLE*. Carl Hanser Verlag GmbH & Co. KG, München. <https://doi.org/10.3139/9783446431164>
- [22] Taavi T. Tajala, Martijn C. Willemsen, and Joseph A. Konstan. 2018. Movie-explorer. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing - SAC '18*. ACM Press, New York, New York, USA, 1383–1392. <https://doi.org/10.1145/3167132.3167281>
- [23] Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. 2013. Escape the Bubble: Guided Exploration of Music Preferences for Serendipity and Novelty. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) (RecSys '13). Association for Computing Machinery, New York, NY, USA, 335–338. <https://doi.org/10.1145/2507157.2507223>
- [24] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- [25] Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A Personalized System for Conversational Recommendations. *J. Artif. Int. Res.* 21, 1 (March 2004), 393–428.
- [26] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the unknown. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing - SAC '18*. ACM Press, New York, New York, USA, 1396–1399. <https://doi.org/10.1145/3167132.3167419>
- [27] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. 2010. Temporal recommendation on graphs via long- and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. ACM Press, New York, New York, USA, 723. <https://doi.org/10.1145/1835804.1835896>